

The effect of improvement of datasets on accuracy achievement in deep learning: an example of disease detection in hops plant

Haluk Tanrikulu^{1,*}, Murat Hüsnü Sazlı² and Hasan Parça³

¹Middle East Technical University, Continuing Education Center, Turkey; ²Ankara University, Department of Electrical and Electronics Engineering, Turkey; ³Ankara University, Department of Agricultural Economics, Turkey

*Corresponding author's e-mail: tanrikul@metu.edu.tr

Plant diseases are a major threat to food safety and security. Preventing the loss of money and time is possible with early diagnosis of plant diseases. Recent advances in computer vision have led to successful methods for the early detection of plant diseases. In this research, images of downy mildew (*Pseudoperonospora humuli*) and powdery mildew (*Podosphaera macularis*) diseases of hops (*Humulus lupulus* - hops) plant were collected over the internet and classified with the most successful Convolutional Neural Network (CNN) model. In order to increase the performance of the CNN model, images that do not contribute to learning were removed from the datasets, and optimum datasets were created by adding new images that comply with the rules we determined. The model was trained with a small number of selected images and detected downy mildew and powdery mildew diseases of hops with high performance. In this study, certain rules were determined in the recognition of plant diseases, the collection of diseased leaf images and the creation of the data set. It has been shown that training datasets created by following these rules increase performance in learning.

Keywords: CNN, deep learning, dataset optimization, few image data.

INTRODUCTION

The early stage of the plant disease first appears on its leaves, as reported in numerous research publications and reports (Kaur 2018). Today, diseased leaf images are taken with smart phones with high resolution cameras and shared on the internet. Researchers create datasets of plant diseases by collecting these images from the internet and share them on data sharing sites such as Kaggle.

However, the number of these diseased plant pictures collected from the internet is not enough to create the necessary training sets for deep learning. Therefore, for data augmentation, methods such as Generative Adversarial Networks (GAN) or geometrically images shifting, rotating, cutting are used (Gomaa 2021). Augmented datasets created by both GAN and other methods contain images that are a kind of copies of the original images. Although learning with data augmentation provides the accuracy rate of learning with original pictures, the process takes time.

The most widely used algorithm for the diagnosis of plant diseases is CNN algorithm. In the study of Rezende *et al.*, the weights of the CNN-based VGG16 and VGG19 models were used to detect 20 different plant diseases from 10 different plant species (Rezende *et al.*, 2019). Abade *et al.*, showed that

the multi-channel CNN (M-CNN) model achieved better accuracies than the transfer learning model (Abade *et al.*, 2019). Uğuz and Uysal showed that the VGG16 model was more successful than the CNN model they suggested in detecting olive plant disease (Uğuz *et al.*, 2021). In the study of Brahimi *et al.*, who used their own CNN model, the images of 14,828 tomato leaves with nine different diseases were used and the success of disease detection reaching 99.18% accuracy was achieved (Brahimi *et al.*, 2017). Ensari *et al.*, created a CNN model that found three different plant diseases from the datasets of maize and grapes and achieved 97.03% accuracy in the tests (Ensari *et al.*, 2020). In the CNN model proposed by Todo and Okura, some layers were optimized and the created attention maps were interpreted and the layers that did not contribute to the result were removed. In this way, it has been shown that the number of parameters can be reduced by 75% without affecting the classification accuracy (Todo *et al.*, 2019). Jiang *et al.*, created a new dataset for apple leaf disease by combining complex images taken from real field conditions and images obtained in the laboratory environment and used it in disease recognition in a CNN model (Jiang *et al.*, 2019). Maheshwari and Shrivastava trained the CNN model they created with the diseased mango leaf training set, which they multiplied with

the data augmentation method. They showed that data augmentation improves the current accuracy performance ratio. They showed that data augmentation increases the current accuracy performance rate (Maheshwari *et al.*, 2020). There are studies in the literature showing that the success of the model will be low if the training and test data sets are not collected under the same conditions. Most of the images in the PlantVillage dataset used in most studies were created under laboratory conditions (Ferentinos, 2018; Mohanty *et al.*, 2016). This means that many factors such as angle of capture, background, symptom area size and light conditions are controlled. Ferentinos trained one model with images taken under controlled laboratory conditions and another model for the same disease with images taken under field conditions. It has been shown that the model trained on images taken under field conditions has better performance in tests with test datasets taken from the field (Ferentinos, 2018). Mohanty *et al.*, and Ferentinos observed that although deep neural network models trained using the PlantVillage dataset achieved classification accuracy in excess of 90%, the accuracy dropped significantly in tests with images outside the PlantVillage dataset and images taken under different conditions. In the study of Mohanty *et al.*, they showed that 90% accuracy rates dropped to just over 31% (Muthukumarana *et al.*, 2020). For this reason, users need to know which environment and under conditions was taken images. It is important to know that users are farmers and they will take images in field conditions with mobile applications. For this reason, it is important for the creators of the model to collect the training images from the field (Mohanty *et al.*, 2016).

Barbedo *et al.*, He states that the failure that Mohanty noted in his study was due to the small number of images collected that did not include symptom regions. Therefore, the model works well on test images taken under the conditions in which images are captured from the training dataset. However, the model performs poorly when tested on images taken on different days, in different locations, and under different capture conditions. Therefore, in order to make progress in the field, it is important to discover what these conditions are and to investigate how they can be reduced (Barbedo *et al.*, 2018).

In this study, we propose a *dataset optimization method* to improve the performance of datasets when the number of images is low. After each failed classification of the created CNN Model, the data set used was carefully analyzed and the reasons for the failure were revealed. The cause of each failure was associated with an image defect or deficiency. Images containing these defects were excluded from the dataset. Trials continued until the increase in accuracy rate stopped. A guideline has been established to avoid such defects in images. The purpose of this article is to demonstrate the increase in performance with the implementation of these

guidelines, which show how to select images that are better suited to the dataset rather than data augmentation.

MATERIALS AND METHODS

Images containing downy mildew and powdery mildew diseases of hops were collected automatically on the internet and these images were examined and labeled one by one with the contributions of expert agricultural engineers. The images of two diseases were trained with the CNN model we created, and the trained model was then tested with the test dataset. In the material part of our study, the hops plant and its diseases are explained, and in the method part, how the optimization of the plant diseases dataset is done with the CNN model is explained.

Material: Hops is a herbaceous plant species that is a member of the cannabaceae family. Most of them are green in color, but when they bloom, they have a white color (Figure 1) (Gent 2010). Its shape resembles a pubescent yellowish greenish cone. It is mostly grown around Bilecik in Turkey.



Figure 1. Hops Plants (Gent 2010).



Figure 2. a) Hops sprouts with downy mildew pathogen b) Powdery mildew disease on a leaf (Gent 2010).

Downy mildew in hops is caused by the fungus-like organism *Pseudoperonospora humuli*. It is one of the most important hop diseases in the Pacific Northwest and worldwide. In addition to yield and quality losses due to downy mildew, it can cause 100% product loss depending on the type of infection (Fig.2) (Gent, 2010).

Powdery mildew disease is caused by the fungus *Podosphaera macularis* and is one of the most important diseases of hops. The disease can cause severe crop damage, in some cases resulting in complete loss of marketable yields due to loss of production and poor cone quality (Gent, 2010).

Method

Creation of datasets: The data sets to be used in learning were collected from the internet and labeled by taking expert

opinion, images of two different diseases of the hops plant (downy mildew and powdery mildew). The collected pictures consist of images taken both in the field and in the laboratory environment. The images produced in the laboratory environment focused directly on the leaf and the disease. Although 1000 pictures of hops disease were downloaded from the internet, only 80 of them were added to the data set as deemed appropriate by the experts. Figure 3 and Figure 4 show some of the images representing the two diseases in the training set.

Dataset optimization: The first dataset collected from the internet and labeled by experts was used to train the CNN model. The accuracy rates of the first training set in two different test sets are 84.21% and 88.89%, respectively. The



Figure 3. Some of the pictures in downy mildew training set



Figure 4. Some of the pictures in the powdery mildew training set

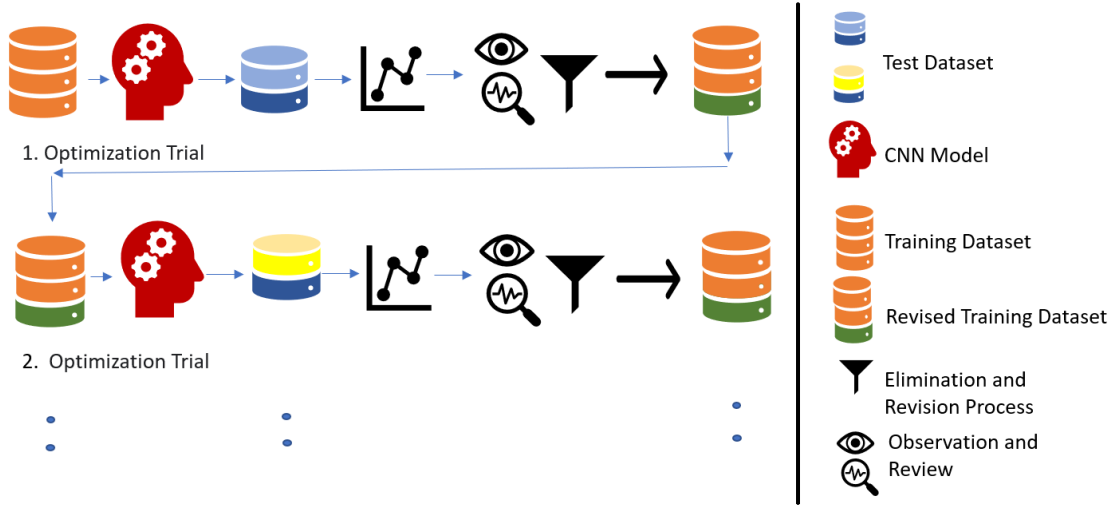


Figure 5. Proposed Data Set Improvement Model

network structure in the CNN model, which is explained in detail in Section 2.3, was used unchanged in subsequent trials. Figure 5 shows how the training set was revised. In the first trial, the training data set consists of 87 downy mildew and 64 powdery mildew diseased images. The validation dataset contains images of 17 downy mildew and 7 powdery mildew diseases.

After the first try, 8 powdery mildew disease images were removed from the original dataset that could reduce performance. The CNN model trained with the revised training set was tested with the test 2 dataset and the accuracy rate was found to be 100%.

In the second trial, revisions were made to both the training and validation datasets. Nine downy mildew images were extracted from the training dataset used in the first trial. Thus, a new training set was created from 76 downy mildew and 64 powdery mildew disease bars. In this trial, the validation data set was created from images showing 10 downy mildew and 6 powdery mildew diseases. It was observed that the disease recognition accuracy rate of the model decreased after the revision in both sets.

Images that reduce the accuracy rate and make learning difficult were removed from the data set and new images were added instead. This study was repeated with new data sets created by following the rules specified in the directive below.

- Because very small resolution (size) pictures are ineffective in learning, removed from the training set.
- Pictures taken remotely were removed from the training set.
- If the images of the diseased lesions remained in a very small area for the whole image, this area was revised and enlarged.
- In the images where many leaves are together, the infected leaves have been resized to stand out.

- Blurred, mixed, crowded background images have been removed.
- Images containing different diseases were excluded from the data set.
- Pictures suitable for these conditions were found by searching the internet again and added to the data set.

At the end of the 5th trial, the model was trained with a dataset containing 77 downy mildew and 71 powdery mildew diseases, and the accuracy rate in tests with test-1 and test-reached 92.21% and 100%, respectively. 2 datasets. This successful model was tried again with two different mixed test data sets collected from the internet in accordance with the rules in the above directive. In these trials, 100% accuracy was obtained in both test sets.

Applications and creation of the model: An application was prepared in the Python programming language (Python) to collect images from the Internet. Python is an object-oriented, interpretive and high-level programming language. The prepared application runs Google's search engine in the background, finds images and creates a folder named with keywords and saves them to the computer. Afterwards, the images deemed appropriate were reclassified and labeled by the experts according to their diseases.

The binary classification model was created using Tensorflow and Keras modules. TensorFlow is a free and open source Python library for machine learning. Similarly, Keras is an open source neural network library written in Python that uses the TensorFlow module in the background.

The summary of the model is given in Figure 6. This successful model was used in all trials. The convolution layers have a normalization layer followed by a pooling layer, and all layers in the network have ReLu nonlinear activation units. The network of our CNN model consists of 3 convolution layers, then a fully connected layer, and finally a

sigmoid layer. After each convolutional layer, there is a pooling layer.

| Layer (type) | Output Shape | Param # |
|--------------------------------|----------------------|---------|
| conv2d (Conv2D) | (None, 148, 148, 16) | 448 |
| max_pooling2d (MaxPooling2D) | (None, 74, 74, 16) | 0 |
| conv2d_1 (Conv2D) | (None, 72, 72, 32) | 4640 |
| max_pooling2d_1 (MaxPooling2D) | (None, 36, 36, 32) | 0 |
| conv2d_2 (Conv2D) | (None, 34, 34, 64) | 18496 |
| max_pooling2d_2 (MaxPooling2D) | (None, 17, 17, 64) | 0 |
| flatten (Flatten) | (None, 18496) | 0 |
| dense (Dense) | (None, 512) | 9470464 |
| dense_1 (Dense) | (None, 1) | 513 |
| Total params: 9,494,561 | | |
| Trainable params: 9,494,561 | | |
| Non-trainable params: 0 | | |

Figure 6. Summary of the model.

RESULTS

Images for many datasets can be found by searching the Internet. However, the relevance of the images found is often unreliable. To confirm the accuracy of diseases in the images collected, agronomists must do meticulous work and label all images with appropriate disease information. As it is known, it is important to use correctly classified and labeled images for the training and validation dataset. In this way, a suitable and reliable detection model can be developed.

In our studies, training and validation sets were created from a small number of hop plant images and the network was

trained on a successful CNN model. In order to increase the accuracy rate of the initial training set, the images were examined by experts and the model's responses to each labeled image in the test were examined. The images that would decrease the performance were removed from the training set, and new training data sets were created and retests were made. As a result, the accuracy rate increased by 8.77% and 11.11% at the end of the 5th trial (Table 1).

According to this directive, two mixed tests were created by collecting hops images from different websites. It was observed that the accuracy rate reached 100% in the trials with mixed test data sets (Table 2).

DISCUSSION

In the literature, it is seen that CNN models trained with datasets with a large number of images provide high performance (Brahimi *et al.*, 2017). In this study, a selected small number of diseased plant images (less than 100 per class) were trained in a deep convolutional neural network. As Mohanty *et al.*, stated, in studies where Planet Village datasets, which are widely used in disease diagnosis, are used, creating the test dataset from different environments (laboratory or field) reduces the success rate.

In 2016, Barbedo classified 1383 diseased images of 12 products obtained from different environments using deep learning and transfer learning methods. Despite the variation in plant species, diseases, and image capture conditions, the study was successful at a rate of 86%.

Literature studies predict that performance will increase with the use of images produced under the same conditions. These conditions include both the physical conditions where the image was taken and the distance limitations from which the image will be obtained. The most important output of this

Table 1. Improvements made on the dataset.

| Dataset | Tests with revised training datasets | | | | | |
|--------------------------|--------------------------------------|----------------|--------------------------|----------------|------------|--------|
| | # of Training examples | | # of validation examples | | Accuracy % | |
| | Downy mildew | Powdery mildew | Downy mildew | Powdery mildew | Test 1 | Test 2 |
| Original Initial Dataset | 87 | 64 | 17 | 7 | 84.21 | 88.89 |
| 1.Experiment 1.Dataset | 87 | 56 | 17 | 7 | 84.21 | 100.00 |
| 2.Experiment 2.Dataset | 76 | 56 | 10 | 6 | 76.92 | 95.65 |
| 3.Experiment 3.Dataset | 77 | 73 | 10 | 6 | 76.92 | 86.36 |
| 4.Experiment 4.Dataset | 77 | 73 | 10 | 6 | 76.62 | 100.00 |
| 5.Experiment 5.Dataset | 77 | 71 | 10 | 6 | 92.31 | 100.00 |

Table 2. Accuracy rates of the model on mixed test datasets.

| Dataset | Tests with revised training datasets | | | | | |
|-------------------------|--------------------------------------|----------------|--------------------------|----------------|---------------------------|---------------------------|
| | # of Training examples | | # of validation examples | | Accuracy % | |
| | Downy mildew | Powdery mildew | Downy mildew | Powdery mildew | Mixed test 1 35 pieces | Mixed test 2 25 pieces |
| 6.Experiment: 5.Dataset | 77 | 71 | 10 | 6 | 100 | 95.45 |
| 7.Experiment: 5.Dataset | 77 | 71 | 10 | 6 | 100 | 100 |

study is that there is a user guide for generating data sets. This guideline should be used in the selection of both the training set and the test set.

Our study was tested multiple times by changing the dataset to select optimized images within a small number of images. In these tests, the reasons that reduced the performance of the images were determined and similar images that would create these reasons were excluded from the training set. A directive was formed by making these determinations a rule. The created directive is given in Table 3.

Table 3. Implemented directives and actions taken.

| Rule | Transactions Taken |
|---|--|
| Extract very small resolution images from dataset | Since the input images feeding the convolution layer are 150X150 pixels, no images below this ratio were used. |
| Extract distant images from the dataset, always use close-up images | Images taken from a distance of 20-60 cm were used. |
| Remove multi-leaf images, use single-leaf images | Only single-leaf images are used in the datasets. Images with many diseased leaves together were not used. |
| If the disease lesions are seen in a very small area, enlarge and revise these parts of the images. | Enlargement or cutting was done to make the lesion stand out. |
| Resize diseased plant leaf to cover at least 25% of the entire image | The diseased leaf was brought forward in the image. |
| Remove blurry, mixed, crowded background images from the dataset | Removed from dataset |
| Extract pictures containing the two diseases together from the dataset. | Extracted from dataset |
| Extract images containing different diseases from the dataset | Extracted from dataset |
| Find pictures of diseased hops on the internet that meet these guidelines and add them to the dataset after expert opinion. | Research was done again. Images conforming to the directive added to the dataset. |

With the study, it has been shown that taking or obtaining the images that make up the data set in a way that meets the same or similar conditions increases the performance. With the use of the directive, certain rules were followed in the selection of the images that make up the training dataset, and high accuracy was achieved with a small number of images.

Conclusion: In this study, it has been shown that while developing deep learning applications with a small number of data sets, determining and applying a directive in accordance with certain rules in image selection increases the accuracy of the model.

Authors Contributions statement: HT conducted the research, analyzed it and created the necessary software and

wrote the paper for his PhD. MHS analyzed and supervised the entire study. HP provided consultancy in the creation of the data set by detecting the diseases.

Conflict of interest: The authors have declared no conflicts of interest for this article

REFERENCES

- Abade, A. S., A. Almeida and F. Vidal. 2019. Plant diseases Recognition from Digital Images using Multichannel Convolutional Neural Networks. In: Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications. VISIGRAPP. pp. 450-458.
- Barbedo, J. G. A. 2018. Factors influencing the use of deep learning for plant disease recognition. Biosystems Engineering. 172:84-91.
- Barbedo, J. G. A. 2018. Impact of dataset size and variety on the effectiveness of deep learning and transfer learning for plant disease classification. Computer Electronics in Agriculture. 153:46-53.
- Brahimi, M., K. Boukhalifa, A. Moussaoui. 2017. Deep Learning for Tomato Diseases: Classification and Symptoms Visualization. Applied Artificial Intelligence.3:299-315.
- Ensari, T., C. D. Armah, A. E. Balsever, M. Dagtekin. 2020. Convolutional Neural Networks for ImageBased Digital Plant Phenotyping. European Journal of Science and Technology. Special Issue:338-342.
- Ferentinos, K. P. 2018. Deep learning models for plant disease detection and diagnosis. Comput. Electron. Agriculture.145:311-318.
- Gent, D. H. 2010. Field Guide for Integrated Pest Management in Hops. US Department of Agriculture, Washington.
- Gomaa, A. A. 2021. Early Prediction of Plant Diseases using CNN and GANs. International Journal of Advanced Computer Science and Applications (IJACSA).12/5:514-519.
- Jiang, P., Y. Chen, B. Liu, D. He, C. Liang. 2019. Real-Time Detection of Apple Leaf Diseases Using Deep Learning Approach Based on Improved Convolutional Neural Networks. IEEE Access V.7:59069-59080.
- Kaur, S., S. Pandey and S. Goel. 2019. Plants Disease Identification and Classification through Leaf Images: A Survey. Archives of Computational Methods in Engineering. 26:507-530.
- Maheshwari, K. and A. Shrivastava. 2020. A Review on Mango Leaf Diseases Identification using Convolution Neural Network. International Journal of Scientific Research & Engineering Trends. 6:1399-1403.
- Mohanty, S. P., D. P. Hughes and M. Salathé. 2016. Using deep learning for image-based plant disease detection. Front Plant Science.7:1-10.

- Muthukumarana, P. S. and A. C. Aponso. 2020. A Review on Deep Learning Based Image Classification of Plant Diseases. *International Journal of Computer Theory and Engineering*. 12:118-122.
- Rezende, V., M. Costa, A. Santos and R. C. L. Oliveria. 2019. Image Processing with Convolutional Neural Networks for Classification of Plant Diseases. In: 2019 8th Brazilian Conference on Intelligent Systems (BRACIS) IEEE Xplore.pp. 705-710.
- Toda, Y. and F. Okura. 2019. How Convolutional Neural Networks Diagnose Plant Disease. *Plant Phenomics*. Article ID. 9237136.
- Uğuz, S, N. Uysal. 2021. Classification of olive leaf diseases using deep convolutional neural networks. *Neural Computing and Applications*. 9:4134-4149.